

The Logic Backbone of a Transcription Network

M. Cosentino Lagomarsino,¹ P. Jona,² and B. Bassetti³

¹UMR 168 / Institut Curie, 26 rue d'Ulm 75005 Paris, France*

²Politecnico di Milano, Dip. Fisica, Pza Leonardo Da Vinci 32, 20133 Milano, Italy

³Università degli Studi di Milano, Dip. Fisica, and I.N.F.N. Via Celoria 16, 20133 Milano, Italy[†]

(Dated: February 7, 2008)

A great part of the effort in the study of coarse grained models of transcription networks concentrates on their dynamical features. In this letter, we consider their *equilibrium* properties, showing that the backbone underlying the dynamic descriptions is an optimization problem. It involves N variables, the gene expression levels, and M constraints, the effects of transcriptional regulation. In the case of Boolean variables and constraints, we investigate the structure of the solutions, and derive phase diagrams. Notably, the model exhibits a connectivity transition between a regime of simple gene control, where the input genes control $O(1)$ other genes, to a regime of complex control, where some “core” input genes control $O(N)$ others.

PACS numbers: 87.10+e, 89.75.Fb, 89.75.Hc

Introduction. Identity, response and architecture of a living system are central topics of molecular biology. Presently, they are largely seen as a result of the interplay between a gene repertoire and the regulatory machinery [1, 2]. Gene transcription in mRNA form is an important step in this process. At this level, the regulatory machinery is embodied by the transcription factors, proteins that bind to special sites along DNA and control the activity of RNA polymerase [3, 4] (Fig. 1). This process is referred to as signal integration. Together, the *cis*-regulatory regions establish a set of interdependencies between transcription factors and genes, including other transcription factors: a “transcription network” [1]. Some of such networks of living organisms are now being explored experimentally [2], and show a modularity that has important biological implications [5]. Understanding the gene expression patterns determined by these networks is an enormous challenge. The problem is that transcription networks are fairly large, so that a coarse graining is needed. This fact has many consequences, mainly related to the dynamics. For example, the pioneering approach of Kauffman [8] suggesting a synchronous deterministic update for a Boolean (i.e. on/off) representation of the network is still being debated [9]. Microscopically, it is well accepted that the Gillespie algorithm [6, 7] correctly describes the events of chemical kinetics involved. On the other hand, with a mesoscopic average in time, it is not clear what the emergent time scales might be.

We approach this problem with a model based on two features. Firstly, it focuses, rather than on dynamics, on the compatibility between gene expression patterns and signal integration functions. Simply put, a cell with N genes can express them in exponentially many ways, 2^N in the Boolean representation. However, the cell never explores all the patterns of expression. It only knows clusters of correlated configurations. An elementary example is the λ -phage [3]. Looking at the system, one could observe the states 10, 01 or perhaps 00, but not 11. One can think that in larger systems many configurations are ruled out for the same compatibility reasons. Secondly, the model takes explicitly into account that some genes are essentially “free” from

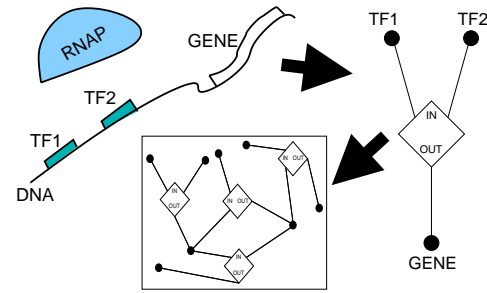


FIG. 1: Schematics of the representation of a transcription network. Each signal integration function at the *cis*-regulatory region of a gene corresponds to a constraint on the gene expression variables. Bottom: example of factor graph of GR1 for $K_b = K = 2$. Each diamond node represents a constraint, while each black circle is a variable.

the point of view of transcription (Fig. 2). This fact is evident looking at the available data [5]. While the biological situation is more complex, we regard these genes as input receptors, connected to external stimuli. The simplest formulation (GR1, from Gene-Regulation) assumes Boolean variables and functions. We use it to investigate theoretically the control exerted by the free genes on the expression patterns for large N , and for random realizations of the constraints [10]. Analysis of the satisfying configurations leads to the introduction of a “core” of network variables. Depending on the number of free genes in the core and the connectivity of the constraints, there are three distinct regimes of gene control. In the first regime, the core is empty. Each free gene controls the state of a small number of genes (“simple control” phase). In the second regime (“complex control”), the core contains free genes that control, both directly and indirectly, order N others. Thus, in the complex control phase, the free core genes can be interpreted as the subset of genes that determine a choice of an expression program. In the third regime, there are no free genes in the core, and the system cannot control the simultaneous expression of all its genes. The transition can be tuned by varying the connectivity and the number of constraints.

Model. The two main ingredients of our representation of a transcription network are: (i) A set of N discrete variables $\{x_i\}_{i=1..N}$ associated to genes or operons (identified with their transcripts and protein products). These variables represent the expression levels. (ii) A set of M interactions, representing the signal integration, $\{I_b(x_{b_0}, x_{b_1}, \dots, x_{b_{K_b}})\}_{b=1..M}$, with $\gamma = M/N \leq 1$. It is useful to represent variables and interactions in a so-called factor graph, as illustrated in Fig. 1. Note that the constraints I_b contain the topology of the graph. The x_i represent real or coarse-grained expression levels and can take values in $\{0, \dots, q\}$ or even continuous ones. In general, using the Shea and Ackers model of gene activation by recruitment [4, 13], one can construct a local free energy associated to each signal integration node, that generates the constraints [14]. Here we consider the simplest possible scenario, GR1, treating the expression levels as Boolean variables (i.e. setting $q = 1$), and the signal integration functions as Boolean functions $\{f_b(x_{i(b,1)}, \dots, x_{i(b,K_b)})\}_{b=1..M}$. The coordinates $i(b, l)$ point at the variable occupying place l in the b th constraint. The expression

$$x_{i(b,0)} = f_b(x_{i(b,1)}, \dots, x_{i(b,K_b)}), \quad (1)$$

imposes that the variable $x_{i(b,0)}$ is the output of the function f_b . For example, let us consider a graph with three variables and one constraint, labeled by $b = 1$. Supposing the first two variables regulate the third, $K_1 = 2$, $i(1,0) = 3$; $i(1,1) = 1$; $i(1,2) = 2$. If, for instance, the transcription can occur only when both regulators are present, then $f_1(x_1, x_2) = 1$ only if $x_1, x_2 = 1$ (Boolean AND function). The local connectivity of a function node is $k_b = 1 + K_b$. K_b is called the “in-degree”. The factor graph is also associated to an “out-degree” $C_i \equiv c_i - 1$, where c_i is the number of functions connected to x_i . The fact that variables and constraints are Boolean make GR1 an optimization problem of the satisfiability type (Sat) [10]. A very special one, given the structure of the signal integration functions. The properties of the model depend on the class of graphs and Boolean functions used. The results presented in this work hold for a rather large class of Boolean functions (see appendix A).

Structure of the solution space. The phase-space structure has never been explored for this particular case. We set out to analyze the number of compatible configurations \mathcal{N} , for large N and M and random instances of the problem. To this end, consider the following argument, which focuses on the control exerted by the $N - M$ “free” input genes on the compatible solutions. Together with the free genes, the network contains some genes which are regulated but do not regulate any other. We can refer to them as “leaves” (Fig 2). Given a realization, a leaf can always be adjusted to the output value of its function, which is then satisfied. Let us now remove from the graph each leaf and iterate this procedure (a variation of the so-called “leaf removal” algorithm [12]). There are two possible outcomes: either erasing all the graph, leaving the free genes as isolated points, or stopping at a core of constraints that contains loops. In other words, the algorithm identifies the tree-like components of the graph connected to outputs.

The core will be composed of N_C genes and M_C constraints. Let us now imagine to have a compatible configuration, flip a free gene, and try to construct another compatible configuration. In the case where the core is empty, since the graph is tree-like, it will always be possible to perform this operation by local rearrangements, which propagate the output of the functions. Thus, flipping all the free genes, we find 2^{N-M} satisfying configurations. In the presence of a core, because of the loops, flipping a free gene of the core will in general rearrange all the core genes, and is not guaranteed to lead to a new satisfying state. In fact, it is not granted there will be a compatible state to start with. Provided there is, the output propagation procedure can be applied to the non-core free genes to construct another. Thus, in general the $N - M$ de-

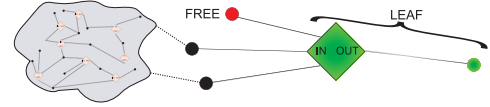


FIG. 2: Example of a leaf. The free gene that regulates it will not be in the core. The other two transcription factors are connected to the rest of the network, represented by a cartoon.

grees of freedom given by the free genes cannot guarantee a solution. The relevant parameter is the number of core free genes $\Delta_C = N_C - M_C$. Let us for the moment restrict to the case of fixed in-degree (“ K -GR1”). In appendix B, we show that the average of \mathcal{N} on the class of all Boolean functions is 2^{N-M} . Thus: (a) if the core is empty, the number of compatible configurations constructed by flipping the free genes are on average all the possible ones. (b) If the core is not empty, in the average case it will still be possible to construct 2^{N-M} solutions by flipping the free genes. If $\Delta_C > 0$, there will be 2^{Δ_C} clusters of solutions, and (c) in the case where the core contains no free genes there will be generally contradictions. We can thus distinguish the three regimes: (a) simple control, (b) complex control, (c) no control. Considering ensembles of random graphs, the regimes above depend on the value of $N_C(N)$, $M_C(M)$, so that a proper order parameter to adopt is $\gamma = M/N$ [10]. The phase diagram can be explored studying the rank and the kernel of the connectivity matrix in the ensemble.

Example: the case of Poisson-distributed c_i . This is the simplest ensemble to consider, where $p(c) = \frac{(k\gamma)^c}{c!} e^{-k\gamma}$ [12]. This probability distribution doesn’t exclude that genes may appear in the functions that regulate them, leaving some freedom of choice. In the simplest case, one finds $N_C(\gamma) = N(m - k\gamma(1 - m)m^{k-1})$ and $M_C(\gamma) = N(\gamma m^k)$, where m is defined by the relation $m(k) + e^{-k\gamma m(k)^{k-1}} - 1 = 0$. This gives the phase diagram as a function of γ . For $\gamma < \gamma_d$ there is simple control. For $\gamma_d < \gamma < \gamma_c$ complex control, and for $\gamma > \gamma_c$ no control. For example, for 4-GR1, $\gamma_d \simeq 0.776$ and $\gamma_c \simeq 0.977$. The regimes of gene control correspond to thermodynamic phases, commonly referred to as SAT, HARD-SAT, and UNSAT phase respectively [11]. Furthermore, it is possible to show rigorously the clustering of

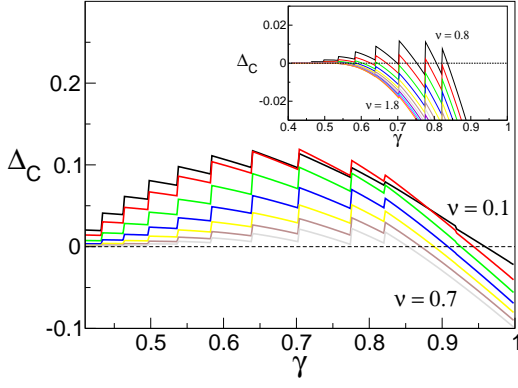


FIG. 3: Δ_C as a function of γ for different values of ν in the multi-Poisson case. The discrete jumps are due to the onset of complex control phases for the different values of k . Δ_C can become negative many times, giving rise to reentrant no control phases (inset). The figure refers to a connectivity distribution with a cutoff at $k = 12$.

solutions argued above. In the simple control phase, one cluster contains all the solutions, and a free gene controls $O(1)$ other genes. The reason for this is that, for Poisson distributed out-degree, the average number of controlled genes is finite ($c\gamma$), while the number of free genes is extensive. Conversely, in the complex control phase, the free genes within the core control $O(N)$ other genes (while there is still $O(1)$ control outside of the core). From a physical point of view, the clusters are separated by an extensive distance, i.e. by free energy barriers. The number of clusters is related to the (computational) complexity Σ of the system, defined by the relation $\log \mathcal{N} \sim N(\Sigma + S)$. Here S , the entropy, measures the width of each cluster, while Σ “counts” the number of clusters. Therefore, by definition, Σ is directly related to Δ_C , i.e. to the partitioning of the free genes in and out of the core. How the system explores (or not) these clusters depends on details of its dynamics. Generically, one can say that the dynamics in a cluster will be residual: many variables are fixed, the others can change. This matches a qualitative feature of many cells, where some genes are constantly expressed, and the rest vary [15].

Multi-Poisson phase diagram. While the fixed k case is useful to get some theoretical insight, the known transcription networks are far from having fixed in-degree. For example, in *E. coli*, the in-degree has Poisson distribution, while the out-degree resembles a power law. For this reason, a biologically more interesting case is when both the in- and the out-degree vary along the network. Considering $p(k|c) = \frac{(k\gamma)^c}{c!} e^{-(k\gamma)}$, the conditioned probability that a variable is in c clauses of the k kind, we have $p(c) = \sum_k \frac{(k\gamma)^c}{c!} e^{-(k\gamma)} p(k)$. The leaf removal equations can be applied separately to sets of clauses with a given connectivity, defining $N_C \equiv \langle N_C \rangle_k$ and $M_C \equiv \langle M_C \rangle_k$, where $\langle X \rangle_k = \sum_p p(k) X(k)$. Choosing $p(k) = Z^{-1}(\nu) e^{-\nu k}$ does not affect the exponential asymptotic decay of $p(c)$ for large c . We can call this case multi-Poisson, as the graph is a superposition, follow-

ing a Poisson distribution, of graphs with fixed in-degree and Poisson-distributed out-degree. The behavior of GR1 on such

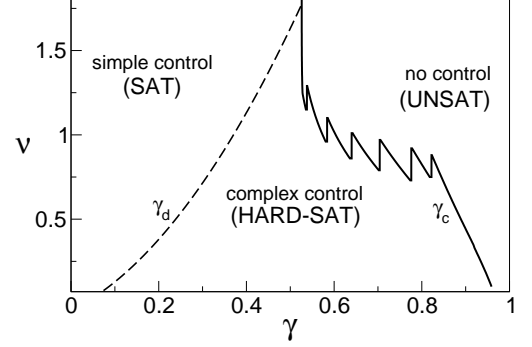


FIG. 4: Phase diagram $\gamma - \nu$ for the multi-Poisson case. The dashed line, a power law with exponent $\zeta \simeq 1.558$, represents the mean value of the numerically evaluated critical parameter $\gamma_d(\nu)$ for the simple-complex control transition of network realizations with $N = 3 \times 10^4$.

a topology is nontrivially different from the fixed connectivity case. The main reason for this is that, while $\Delta_C(\gamma)$ is still locally decreasing, many new discontinuities emerge, due to the influence of clauses with different connectivities. This gives rise to different phenomena. Firstly, Δ_C can increase globally with increasing γ . Indeed, it does increase step-wise with γ after γ_d , to decrease again before γ_c . Its discrete jumps are due to the onset of complex control phases for the different values of k (Fig. 3). This fact has an influence on the number of compatible states as a function of γ . Secondly, Δ_C can become negative and then jump back to a positive value, creating a reentrant UNSAT phase (Fig. 4). This means that, on average, by increasing the number of constraints one can pass from unsolvable problems to solvable ones. A heuristic explanation for this counterintuitive fact is that, at fixed N , the addition of a constraint might connect a closed loop in the core to external free genes, thereby solving a contradiction. Interestingly, the simple to complex control boundary is a power-law in γ (Fig. 4).

Discussion. To conclude, we established a simple framework for the modeling of large scale transcription networks. It is a compatibility analysis on the constraints established by transcription. Its advantage is that, while avoiding to deal directly with the dynamics, it gives non-trivial results. In the absence of an explicit knowledge of the emergent time scales, we feel this is an appropriate approach, particularly in the Boolean approximation treated here. From a technical standpoint, GR1 is different from other problems of the satisfiability kind because of the particular structure of its constraints. This makes it possible to apply the leaf removal technique, which is ineffective for other models, such as random-k-Sat [11]. From a general standpoint, our model shows that the “biological” complexity is not simply measured by the number of genes. A more proper indicator is Δ_C which depends on the order parameter ν , or - roughly - on the number of transcription factors per gene. At fixed number of genes,

it is known that this quantity increases in bacteria that need to react to more environments [17]. Imagining that prokaryotes are naturally found in a simple control phase, our phase diagram predicts an intrinsic limit to this process, represented by the phase boundary with the complex control phase. The multi-Poisson case gives an interesting prediction for the behaviour of this boundary at fixed N . Namely, at criticality, the average number of constraints scales as a power-law with γ . This feature, together with the existence of a core and the predictions on the control exerted by free genes, can possibly be tested experimentally.

The approach presented here is new, and largely unexplored. It is naturally fit to study networks with non-Boolean variables and probabilistic constraints. It can be of use for models that describe other regulation mechanisms than just transcription. More far-reaching extensions include evolutionary models. It is not clear yet exactly how useful it can be for the study of concrete biological networks. Together with the general trends, a biologically significant model has to be able to deal with the details of an individual realization the system. This is, we think, the main challenge to our approach, and the direction we are currently exploring.

We would like to acknowledge interesting discussions with J. Berg, M. Caselle, L. Finzi, M. Leone, A. Sportiello, P.R. Tenwolde, R. Zecchina. We thank an anonymous referee for help improving our manuscript.

APPENDIX

A. GR1 as a Satisfiability problem In this appendix we show how a realization of GR1 can be formulated as a satisfiability problem (Sat) [10], an optimization problem where N Boolean variables are constrained by O conjunctive normal form (CNF) constraints (i.e. by a Boolean polynomial constructed as a product (\wedge) of O disjunctive monomials (\vee)). Equation (1) is equivalent to the XOR Boolean constraint $I_b = \neg(x_{i(b,0)} \vee f_b)$. This can be recast in CNF, as 2^{K_b} clauses involving k_b variables. Each clause corresponds with a simple map to each line $x_{i(b,1)}, \dots, x_{i(b,K_b)}, x_{i(b,0)}$ (including the output) in the truth table of f_b . Namely, if the value of variable $x_{i(b,j)}$ is 1 in the truth table line, it will be negated in the CNF clause. Viceversa, it will be affirmed if its value is 0. The formula $I = \bigwedge_{b=1..M} I_b$, defines a Satisfiability problem on the variables x_1, \dots, x_N , with $O = \sum_{b=1..M} 2^{K_b}$. A realization of GR1 differs from a Sat problem for the intrinsically asymmetric form of the constraints, which “force” the value of $x_{i(b,0)}$. Moreover, considering random instances of the problem, the space of allowed functions of GR1 is much smaller than the corresponding Sat problem. For example, for a clause with fixed connectivity k , while there are 2^{2^k} possible Boolean functions, all of which can appear in Sat, only 2^{2^k} of these can appear in GR1. The two above features make the leaf removal technique useful for the latter model.

B. Average Number of Solutions In this appendix, we discuss the average of \mathcal{N} on the realizations of the constraints

$\{\vec{I}, \vec{f}\}$, for K -GR1. One can write

$$\mathcal{N}(\vec{I}, \vec{f}) = \sum_{\vec{x}} \prod_{b=1}^M \delta(x_{i(b,0)}; f_b(x_{i(b,1)}, \dots, x_{i(b,K_b)})).$$

Here, the randomness is contained: (i) in the specification of the network structure, $\vec{I} = (I_1, \dots, I_M)$, i.e. in the coordinates $i(b, l)$; (ii) in the specification of the functions $\vec{f} = (f_1, \dots, f_M)$ within the class \mathcal{F} of all Boolean functions. An overbar ($\bar{}$) indicates an average on both distributions, $p(\vec{I})$ and $p(\vec{f})$. Taking for \mathcal{F} the class of all Boolean functions, it is straightforward to obtain $\bar{\mathcal{N}} = 2^{N-M}$, independently from the specification of the network structure. This result remains true considering a sub-family \mathcal{F}_ρ of functions that satisfy $\frac{1}{2^{2^K}} \sum_{f \in \mathcal{F}_\rho} p(f) f(\vec{x}) = \rho$. The reason for this is that one finds $\bar{\mathcal{N}} = \sum_{\vec{x}, \vec{I}} p(\vec{I}) \prod_{b=1}^M (\rho \delta_{1; x_{i(b,0)}} + (1 - \rho) \delta_{0; x_{i(b,0)}})$.

* e-mail address: mcl@curie.fr

† e-mail address: bassetti@mi.infn.it

- [1] M. Babu, *et al.*, Curr Opin Struct Biol **14**, 283 (2004).
- [2] M. Herrgard, *et al.*, Curr Opin Biotechnol **15**, 70 (2004).
- [3] M. Ptashne, *A Genetic Switch* (Cell Press, MA, 1992).
- [4] N. Buchler, *et al.*, Proc Natl Acad Sci USA **100**, 5136 (2003).
- [5] S. Shen-Orr, *et al.*, Nat Genet **31**, 64 (2002).
- [6] H. McAdams and A. Arkin, Proc Natl Acad Sci USA **94**, 814 (1997).
- [7] D. Gillespie, J. Phys. Chem. **81**, 2340 61 (1977).
- [8] S. Kauffman, *The Origins of Order* (Oxford Univ. Press, New York, 1993).
- [9] C. Gershenson, in *Artificial Life IX Workshops and Tutorials* (2004).
- [10] S. Mertens, Comput Sci and Eng **4**, 31 (2002).
- [11] M. Mezard, *et al.*, Science **297**, 812 (2002).
- [12] M. Mezard, *et al.*, J Stat Phys **505** (2003). See also S. Cocco *et al.*, Phys Rev Lett **90**, 047205 (2003) and L. Correale *et al.*, cond-mat/0412443.
- [13] M. Shea and G. Ackers, J Mol Biol **181**, 211 (1985).
- [14] M. Cosentino Lagomarsino, *et al.*, (2005), q-bio.MN/0502017.
- [15] We should note that the conventional average of \mathcal{N} might be biased by the weight of exceptions [16]. To access the typical behavior of the system, the correct quantity to compute is the “quenched average”, $\overline{\log \mathcal{N}}$, usually accessed with the replica, or similar methods [11]. In the case under exam we have computed the *annealed* average $\log \bar{\mathcal{N}}$ (in general $\log \bar{\mathcal{N}} \leq \overline{\log \mathcal{N}}$). For K-GR1 with Poisson distributed out-degree, we have shown the presence of a self-averaging property for \mathcal{N} . That is, the quantity $(\overline{\mathcal{N}^2} - \bar{\mathcal{N}}^2) / (\bar{\mathcal{N}}^2)$ vanishes in the thermodynamic limit $N \rightarrow \infty$ so that, in the simple and complex control regimes, an equality holds between quenched and annealed average [14].
- [16] M. Mezard, *et al.*, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
- [17] I. Cases, *et al.*, Trends Microbiol **11**, 248 (2003).
- [18] E. van Nimwegen, Trends Genet **19**, 479 (2003).